Predictive Analytics with Big Social Data

Niels Buus Lassen[1], Rene Madsen[1], & Ravi Vatrapu[1,2]
[1]Computational Social Science Laboratory (CSSL)
Department of IT Management, Copenhagen Business School, Denmark
[2]Westerdals Olso School of Arts, Communication and Technology, Norway
nbl@evalua.dk, john.rene.madsen@gmail.com, vatrapu@cbs.dk

Abstract

Recent research in the field of computational social science have shown how data resulting from the widespread adoption and use of social media channels such as twitter can be used to predict outcomes such as movie revenues, election winners, localized moods, and epidemic outbreaks. Underlying assumptions for this research stream on predictive analytics are that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product and that the shared digital artefact that is persistent can create social influence. In this paper, we demonstrate how social media data from twitter and facebook can be used to predict the quarterly sales of iPhones and revenues of H&M respectively. Based on a conceptual model of social data consisting of social graph (actors, actions, activities, and artefacts) and social text (topics, keywords, pronouns, and sentiments), we develop and evaluate linear regression models that transform (a) iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks (Lassen, Madsen, & Vatrapu, 2014)and (b) facebook likes into a prediction of the global revenue of the fast fashion company, H&M. We discuss the findings and conclude with implications for predictive analytics with big social data.

Research Question

Our basic premise is that social media actions can serve as proxies for user's attention and as such have predictive power. Our central research question is: *to what extent can big social data predict real-world outcomes such as sales and revenues?*

Related Work

We deliberately limit the review of extant literature to empirical work that examined the relationship between social data measures (such as facebook posts/likes/comments/shares, and twitter tweets/re-tweets/mentions/polarity etc.) and real-world business outcomes (revenues, stock price etc.). There has been substantial research work (Bakshy, Simmons, Huffaker, Teng, & Adamic, 2010; Bollen & Mao, 2011; Dorr & Denton, 2009; Gavrilov, Anguelov, Indyk, & Motwani, 2000; Kharratzadeh & Coates, 2012; Mittermayer, 2004) in the direction of predicting the stock prices of the companies based on the analysis of content from the online media such as news items, web blogs, twitter feeds. For example, Gavrilov et al., (2000) applied data mining techniques on the stock information from various companies by clustering them according to their Standard and Poor (S&P) 500 index, whereas the content from the weblogs is used by Kharratzadeh & Coates (2012) to identify the underlying relationships between the companies to make predictions about the evolution of stock prices. The most notable paper in this regard is from Asur & Huberman (2010) who showed that social media feeds can be used as effective indicators of the real-world performance. In their work, they used analysis of hourly rate of tweets about movies, their re-tweets and sentiment polarity to accurately forecast the box-office revenues. In fact, their prediction of movie revenues based on the social data measures from twitter outperformed the leading market-based predictions of the Hollywood Stock Exchange. In terms of macro-societal relationships, a research study investigated whether the public mood as measured from large-scale collection of Twitter tweets can be correlated or even predictive of Dow Jones Industrial Average (DJIA) values has been explored by Bollen and Mao (2011).

Method

We adhered to the methodological schematic recommended by Shmueli and Koppius (2011) for building empirical predictive models. We followed Shmueli and Koppius's (2011) eight methodological steps of predictive model building as depicted in Figure 1.

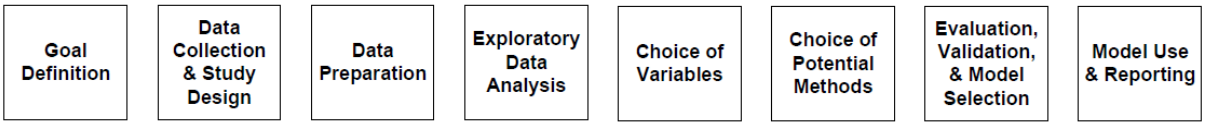

Figure 1: Methodological Steps in Predictive Model Building (Shmueli and Koppius 2011, p. 563)

Dataset

Table 1 below presents the dataset collected for predictive analytics purposes of this paper.

Company	Data Source	Time Period	Size of Dataset

Apple[1]	Twitter	2007 → October 12, 2014	500 million+ tweets containing "iPhone"
H & M[2]	Facebook	January 01, 2009 → October 12, 2014	~15 million Facebook events

Table 1: Overview of Dataset

Predictive Method

We adopt the method of Asur & Huberman (2010) and examine if the same principles for predicting movie revenue with Twitter data can be used to predict iPhone sales and H&M revenues for facebook data. That is, if a tweet/like can serve as a proxy for a user's attention towards a product and an underlying intention to purchase and/or recommend it. We extend Asur and Huberman (2010) in three important ways: (a) addition of facebook social data, (b) theoretically informed time-lagging of the independent variable, social media actions, and (c) domain-specific seasonal weighting of the dependent variable, sales/revenues.

Results

Figures 2 and 3 present the predicted vs. actual sales of iPhone (from tweets contacting the keyword "iphone") and revenues of H&M (from likes on the facebook posts and comments by H&M).

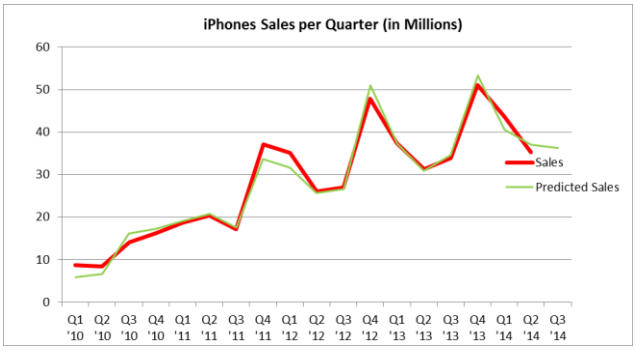

Figure 2: Predicted vs. Actual Sales of iPhones from Tweets

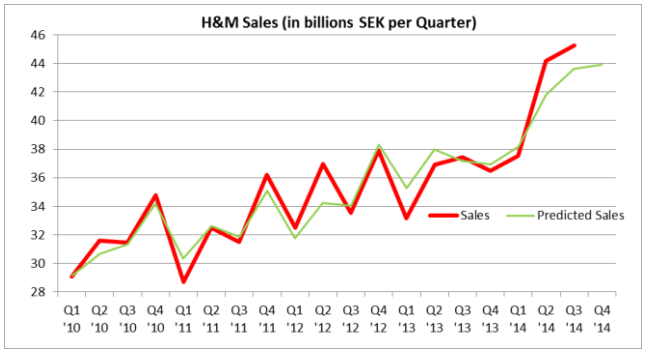

Figure 3: Predicted vs. Actual Revenues of H&M from facebook likes

Conclusions

Drawing from the theoretical framework of AIDA (Awareness, Interest, Desire and Action) and Hierarchy of Effects models in marketing (Belch, Belch, Kerr, & Powell, 2008), combined with the assumption that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product, we demonstrated how social media data from twitter and facebook can be used to predict real-world outcomes such as sales and revenues. Our main contribution is a general purpose prediction model for big social presented in figure 4 below:

$$A_s = A_r \times \frac{U}{C} \times B$$

$$y = \beta_a \times A_r \times \frac{U}{C} \times B + \beta_p \times P + \beta_d \times D + \varepsilon$$

- A_s: *Activity in Social Media (tweets, likes, postings etc)*
- A_r: *Attention in Real World*
- U: *number of Users for the Social Media*
- C: *Potential Customers*
- B: *Media Activity Behavior*
- P: *Presence of Domain-Specific Parameter*
- D: *Distribution of Domain-Specific Parameter*

Figure 4: General Big Social Data Predictive Model

References

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

Bakshy, E., Simmons, M. P., Huffaker, D., Teng, C., & Adamic, L. (2010). The social dynamics of economic activity in a virtual world. *ICWSM2010.* http://misc.si.umich.edu/publications/18.

Belch, G. E., Belch, M. A., Kerr, G. F., & Powell, I. (2008). *Advertising and promotion: An integrated marketing communications perspective*: Mcgraw-Hill.

Bollen, J., & Mao, H. (2011). Twitter mood as a stock market predictor. *Computer*, 91-94.

[1] URL: https://www.apple.com.
[2] URL: http://www.hm.com.

Dorr, D. H., & Denton, A. M. (2009). Establishing relationships among patterns in stock market data. *Data & Knowledge Engineering, 68*(3), 318-337.

Gavrilov, M., Anguelov, D., Indyk, P., & Motwani, R. (2000). *Mining the stock market (extended abstract): which measure is best?* Paper presented at the Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining.

Kharratzadeh, M., & Coates, M. (2012). *Weblog Analysis for Predicting Correlations in Stock Price Evolutions.* Paper presented at the ICWSM.

Lassen, N., Madsen, R., & Vatrapu, R. (2014). Predicting iPhone Sales from iPhone Tweets. *Proceedings of IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC 2014), Ulm, Germany*, 81-90, ISBN: 1541-7719/1514, DOI: 1510.1109/EDOC.2014.1520.

Mittermayer, M.-A. (2004). *Forecasting intraday stock price trends with text mining techniques.* Paper presented at the System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly, 35*(3), 553-572.